

#The key feature of Longitudinal Data Analysis is the analysis of longitudinal
#data, i.e. repeated observations taken on individuals. Oftentimes, the primary
#goal is to characterize the change in response over time (or measurements), and
#factors that influence this change.

#This study was conducted in 16 boys and 11 girls, who at ages 8, 10, 12, and 14
#had their distance (mm) from the center of the pituitary gland to the
#pteryomaxillary
#fissure measured. Changes in pituitary-pteryomaxillary distances during growth is
#important in orthodontal therapy.

#The goals of the study were to describe the distance in boys and girls as simple
#functions of age, and then to compare the functions for boys and girls.

```
library(labelled) # labeling data
library(rstatix)  # summary statistics
library(ggpubr)   # convenient summary statistics and plots
library(GGally)   # advanced plot
library(car)      # useful for anova/wald test
library(Epi)      # easy getting CI for model coef/pred
library(lme4)     # linear mixed-effects models
library(lmerTest) # test for linear mixed-effects models
library(emmeans)  # marginal means
library(multcomp) # CI for linear combinations of model coef
library(geepack)  # generalized estimating equations
library(ggeffects) # marginal effects, adjusted predictions
library(gt)       # nice tables
library(tidyverse) # for everything (data manipulation, visualization, coding, and more)
```

```

library(mvtnorm)

#####

#####

#load("data/dental.RData")
mydata<-read.csv("Dental_Data.csv")

head(mydata)

#Comment: The dental data is in the wide format.

#The longitudinal data must be in long format.
dental_long <- pivot_longer(mydata, cols = starts_with("y"),
                           names_to = "measurement", values_to = "distance") %>%
mutate(
  age = parse_number(measurement),
  measurement = fct_inorder(paste("Measure at age", age))
) %>%
set_variable_labels(
  age = "Age of the child at measurement",
  measurement = "Label for time measurement",
  distance = "Measurement"
)

head(dental_long)

dental_long<-dental_long[,-c(1)]

dental_long<-data.frame(dental_long)

```

```
head(dental_long)
```

```
# Boxplots by measurement
```

```
library(ggplot2)
```

```
ggplot(dental_long, aes(measurement, distance, fill = measurement)) +  
  geom_boxplot() +  
  geom_jitter(width = 0.2) +  
  guides(fill = "none") +  
  labs(x = "", y = "Dental growth, mm")
```

```
# Descriptive Statistics by measurement
```

```
group_by(dental_long, sex, measurement) %>%  
  get_summary_stats(distance, show = c("mean", "sd"))
```

```
#The average response increases both among boys and girls.
```

```
ggplot(dental_long, aes(sex, distance, fill = measurement)) +  
  geom_boxplot() +  
  labs(x = "", y = "Dental growth, mm", fill = "")
```

```
group_by(dental_long, sex, measurement) %>%  
  summarise(mean_distance = mean(distance), .groups = "drop") %>%  
  ggplot(aes(sex, mean_distance, fill = measurement, label =  
    round(mean_distance))) +  
  geom_col(position = "dodge") +  
  geom_text(position = position_dodge(width = 0.9), vjust = -0.5) +  
  coord_flip() +  
  labs(x = "", y = "Mean Dental growth, mm", fill = "")
```

```
#####  
#####
```

```
#Correaltion of the response
```

```
# co-variance matrix
```

```
cov_obs <- select(mydata, starts_with("y")) %>%
```

```
  cov()
```

```
cov_obs
```

```
# correlation matrix
```

```
cov2cor(cov_obs)
```

```
library(GGally)
```

```
ggpairs(select(mydata, starts_with("y")), lower = list(continuous = "smooth"))
```

```
ggpairs(mydata, mapping = aes(colour = sex), columns = 3:6,
```

```
  lower = list(continuous = "smooth"))
```

```
#Trajectory over time
```

```
group_by(dental_long, sex, age) %>%
```

```
  summarise(mean = list(mean_ci(distance)), .groups = "drop") %>%
```

```
  unnest_wider(mean) %>%
```

```
  mutate(agex = age - .05 + .05*(sex == "Boy")) %>%
```

```
  ggplot(aes(agex, y, col = sex, shape = sex)) +
```

```
  geom_point() +
```

```
  geom_errorbar(aes(ymin = ymin, ymax = ymax), width = 0.2) +
```

```
  geom_line() +
```

```
  labs(x = "Age, years", y = "Mean Dental growth, mm", shape = "Sex", col = "Sex")
```

```
#####  
#####
```

#Plot of individual data over time

```
ggplot(dental_long, aes(age, distance, col = factor(id))) +  
  geom_point() +  
  geom_line() +  
  facet_wrap(~ id) +  
  labs(x = "Age, years", y = "Dental growth, mm", col = "Child id") +  
  guides(col = guide_legend(nrow = 3))
```

#Spaghetti plot

```
ggplot(dental_long, aes(age, distance, col = factor(id))) +  
  geom_line() +  
  labs(x = "Age, years", y = "Dental growth, mm", col = "Child id") +  
  guides(col = guide_legend(nrow = 3))
```

#Spaghetti plot by sex

```
ggplot(dental_long, aes(age, distance)) +  
  geom_line(aes(group = factor(id))) +  
  geom_smooth() +  
  facet_grid(~ sex)
```

#Linear Mixed Models

#In these models, some subset of the regression coefficients vary randomly from
#one individual to another, accounting for natural heterogeneity across subjects.
#Individuals in population are assumed to have their own subject-specific mean
#response trajectories over time.

#The mean response is modeled as a combination of population characteristics
#(fixed effects) assumed to be shared by all individuals, while subject-specific
#effects (random effects) are unique to a particular individual.

#Linear Mixed Models are a particular type of hierarchical models which contains
#both fixed and random effects.

#Estimation

#The coefficients can be estimated using Maximum Likelihood (ML) estimation,
#which give estimates that maximize the log-likelihood, i.e. the probability for
#the observed the data given the parameters.

#Maximum Likelihood is not taking account of the degrees of freedom used for
#estimating fixed effects when estimating variance components. This can lead to
#bias estimates (underestimation) for the variance components in data with
#relatively small number of individuals.

#The alternative Restricted Maximum Likelihood (REML) can overcome this
limitation

#by maximizing a marginal or residual (invariant to fixed effects) likelihood.

#Fitting model in R

#Linear mixed-effects models can be fitted using the lmer function in the lme4
#package.

#The first part of the formula model specify the fixed effects part of the model,
#while the second in parantheses specify the random components

```
m1<- lmer(distance ~ 1 + (1 | id), data = dental_long)
```

```
summary(m1)
```

```
ci.lin(m1)
```

```
#The estimated marginal mean of the dental distance is the intercept is 24.02mm
```

```
#The estimated variance of the random-effect reflecting between-subject variability
```

```
#is 3.752; the estimated variance of the error term reflecting within-subject
```

```
#variability is 4.930. The correlation between any two repeated measures (ICC)
```

```
#is equal to  $3.76/(3.76+4.93)=0.43$ 
```

```
#Variance components can be tested using an ANOVA-like table that can be derived
```

```
#using the ranova function in the lmer test
```

```
ranova(m1)
```

```
#The small p-value suggests evidence of between-individual heterogeneity, which
```

```
#support evidence for choosing a mixed-effects model instead of a only fixed-effects
```

```
#model.
```

```
#Graphical presentation of the model
```

```
ggplot(dental_long, aes(id, distance)) +
```

```
  geom_point(aes(col = measurement, shape = measurement)) +
```

```
  geom_point(data = group_by(dental_long, id) %>%
```

```
    summarise(distance = mean(distance), .groups = "drop"),
```

```
    aes(col = "Mean", shape = "Mean"), size = 2.5) +
```

```
  geom_hline(yintercept = mean(dental_long$distance)) +
```

```
  scale_shape_manual(values = c(4, 19, 19, 19, 19)) +
```

```
  labs(x = "Child id", y = "Dental growth, mm", col = "Measurement", shape =
```

"Measurement")

```
#####  
#####
```

#Time effect

#Question: Is the mean response varying with time?

#One modeling strategy is to use indicator variables to contrast mean responses
#at different occasions. In the dental growth example we have measurements of
the

#response at age 8 (baseline), 10, 12, and 14 years-old:

```
m2<- lmer(distance ~ measurement + (1 | id), data = dental_long)  
summary(m2)
```

#The mean dental distance at age 8 years (baseline) is 22.18 mm.

#The difference between the mean responses of 10 and 8 years-old children is mm.

#The difference between the mean responses of 12 and 8 years-old children is mm.

#The difference between the mean responses of 14 and 8 years-old children is mm.

#The emmeans function can be usefull for computing marginal means with
corresponding

#confidence intervals.

```
library(tidyr)
```

```
tidy(emmeans(m2, "measurement"), conf.int = TRUE)
```

#We can test whether the mean response is constant over time by testing the null

#hypothesis that all the regression coefficients used to model time are
simultaneously

#equal to zero


```
Anova(m2)
```

```
#Results from the Wald-test suggest that the trajectory of the mean response over  
#time is not flat.
```

```
#We can presented the estimated trajectory from the fitted model by using the  
#predict function.
```

```
dental_fit <- bind_cols(  
  dental_long, pred_age = predict(m2, re.form = ~ 0)  
)  
ggplot(dental_fit, aes(age, distance)) +  
  geom_line(aes(group = factor(id))) +  
  geom_point(aes(y = pred_age, col = "blue", size = 2) +  
  labs(x = "Age, years", y = "Dental growth, mm")
```

```
#Group effect
```

```
#Question: Is the mean response varying in the two groups of individuals?
```

```
m3<- lmer(distance ~ measurement + sex + (1 | id), data = dental_long)  
summary(m3)
```

```
ci.lin(m3)
```

```
#The difference between mean dental distances of boys vs. girls is 2.3
```

```
 #(95% CI = 0.83-3.81) mm higher than girls, adjusting for time.
```

```
#The emmeans can be used for estimating the mean responses for different  
#combinations of sex and time measurements.
```

```
tidy(emmeans(m3, c("measurement", "sex")), conf.int = TRUE)
```

#A graphical presentation of the fitted model can be easily obtained by

```
dental_fit$pred_agesex <- predict(m3, re.form = ~ 0)
```

```
ggplot(dental_fit, aes(age, distance)) +
```

```
  geom_line(aes(group = factor(id))) +
```

```
  geom_point(aes(y = pred_agesex, col = sex), size = 2) +
```

```
  labs(x = "Age, years", y = "Dental growth, mm", col = "Sex")
```

#Interaction between time and group

#Question: Is the change of the mean response over time varying according to group

#of individuals?

```
m4<- lmer(distance ~ measurement*sex + (1 | id), data = dental_long)
```

```
summary(m4)
```

#The mean distance among girls at age 8 was 21.2 mm

#The mean distance among girls at age 10 was 1.04 mm higher than age 8.

#The mean distance among girls at age 14 was 2.9.0 mm higher than age 8.

#The mean distance among boys at age 8 was $21.2 + 1.7 = 22.9$ mm

#The mean distance among boys at age 10 was $1.04 - 0.11 = 0.93$ mm higher than age 8.

#Testing the interaction between group and time

```
Anova(m4, type = 3)
```

#The Wald test detects some evidence of interaction between age and sex (

= 0.07), which suggests that the mean response profiles might not be parallel.

#The marginal prediction for the combination of sex and time of measurement can
#be obtained as

```
tidy(emmeans(lin_agesexinter, c("measurement", "sex")), conf.int = TRUE)
```

#The difference between the mean response among boys and girls depends on age.
#The change in the mean response over time depends on sex

```
dental_fit$pred_agesexinter <- predict(m4, re.form = ~ 0)
ggplot(dental_fit, aes(age, distance)) +
  geom_line(aes(group = factor(id))) +
  geom_point(aes(y = pred_agesexinter, col = sex), size = 2) +
  labs(x = "Age, years", y = "Dental growth, mm", col = "Sex")
```

```
tidy(emmeans(m4, c("measurement", "sex")), conf.int = TRUE)
```

#Random intercept and slope

#Fitting model in R

```
m5<- lmer(distance ~ age + (age | id), data = dental_long)
summary(m5)
```

#This model assumes that individuals vary not only in their baseline level
#of response (intercept), but also in terms of their changes (slope) in the
#mean response over time.

```
VarCorr(m5)
```

```
as.data.frame(VarCorr(m5))
```

#There is a negative correlation (-0.32) between random intercepts and slopes:

#the greater is the mean response at baseline (8 years), the lower is likely to
#be the increase (slope) of the rate of change over time.

#We can get the predicted value of the marginal (overall) and individual-specific
#mean response trajectories.

```
expand.grid(
  age = seq(8, 14, .5),
  id = unique(dental_long$id)
) %>%
  bind_cols(
    indiv_pred = predict(m5, newdata = .),
    marg_pred = predict(m5, newdata = ., re.form = ~ 0)
  ) %>%
  ggplot(aes(age, indiv_pred, group = id)) +
  geom_line(col = "grey") +
  geom_line(aes(y = marg_pred), col = "blue", lwd = 1.5) +
  labs(x = "Age, years", y = "Dental growth, mm")
```

#Question: Is there any difference between boys and girls?

```
m6<- lmer(distance ~ age*sex + (age | id), data = dental_long)
summary(m6)
```

#There is evidence of significant interaction between sex and time in
#predicting dental distance (p for interaction = 0.032).

```
expand.grid(
  age = seq(8, 14, .5),
  id = unique(dental_long$id)
```

```
) %>%  
left_join(select(mydata, sex, id), by = "id") %>%  
bind_cols(  
  indiv_pred = predict(m6, newdata = .),  
  marg_pred = predict(m6, newdata = ., re.form = ~ 0)  
) %>%  
ggplot(aes(age, indiv_pred, group = id)) +  
geom_line(col = "grey") +  
geom_line(aes(y = marg_pred, col = sex), lwd = 1.5) +  
labs(x = "Age, measurements", y = "Mean Dental growth, mm")
```